MODELING CUSTOMER CHURN AMONG MOVERS IN THE FINNISH INSURANCE MARKET

Eero Einar Sahlberg
013762144
University of Helsinki
Faculty of Social Sciences
Master's Thesis in Economics
January 2018

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty Valtiotieteellinen / Social Sciences		Laitos/Institution– Department Politiikan ja talouden tutkimuksen laitos / Dept. of political and economic studies
Tekijä/Författare – Author Sahlberg, Eero Einar		
Työn nimi / Arbetets titel – Title Modeling customer churn among movers in the Finnish insurance market		
Oppiaine /Läroämne – Subject Taloustiede / Economics		
Työn laji/Arbetets art – Level Pro Gradu –tutkielma / Master's Thesis	Aika/Datum – Month and year 01/2018	Sivumäärä/ Sidoantal – Number of pages 45 (+ 7 in Appendix)

Tiivistelmä/Referat – Abstract

This thesis examines underlying causes of customer churn in the Finnish insurance market. Using individual data on moving insurance customers, econometric modeling is conducted to find significant relations between observed customer characteristics and behavior, and the probability to churn.

A subscription-based business gains revenue not only from new sales but more importantly from automatic renewals of existing customers, i.e. retention. Significant drops in retention are important to understand for the insurer in order to not lose profit. Churn is an antonym for retention.

A change of address – or moving homes – is an event around which churn rates spike, as it is a time when all address-specific subscriptions (electricity, internet, etc.) need to be proactively renewed by the consumer. There were one million moving individuals in 2016, as reported by Posti. This means that a significant share of an insurer's customers are at a heightened risk to churn, with an address change being the common denominator. This thesis asks which customer characteristics and experiences significantly either increase or decrease the probability of a customer either changing their home insurance or churning completely around the time of their move.

Insurance literature such as Hillson & Murray-Webster (2007) and Vaughan (1996) are reviewed to present the nature of risk, the insurance mechanism and the modern insurance business model. An annual report by Finance Finland (2017) provides accounting data via which the Finnish market situation is presented, while data and reports by Posti (2016; 2017a; 2017b) provide the numbers and facts regarding Finnish movers. Churn modeling is based on 20th century discrete choice theory, literature of which is reviewed, most notably by Nobel-laureate Daniel McFadden (1974; 2000). Also presented are modern applications of choice theory into churn problems, such as Madden et al (1999).

The empirical section of the thesis consists of data presentation, model construction and evaluation and finally discussion of the results. The final sample of customer data consists of 24 230 observations with 21 variables. Following Madden et al (1999) and with help from Cox (1958) and McFadden (1974), binomial logistic regression models are constructed to relate the probability of churning with the specified variables.

It is found that customer data can be used to predict churn among movers. Significant weights are found for variables denoting the size of a customer's insurance portfolio as well as customer age and the duration of customership. Also the presence of personal insurance products and contact with one's insurer notably affect retention positively. Younger segments and customers with implications of lower income (with fewer insurance products, more payment installments) exhibit a significantly increased probability of churning.

Avainsanat – Nyckelord – Keywords
Choice theory, risk, logit model, logistic regression, subscription, churn, insurance

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Acknowledgements

Writing these words means that my journey through the Finnish education system is nearing its end – at least for the time being. It took a while. As I finally finish this thesis and turn it in I will, for the first time since 1997, not be enrolled in a school. It ought to be a scary notion, but it is not and for that I have many individuals and institutions to thank.

The Finnish education system has allowed me the freedom, space and time to search and explore. To find a passion to pursue without having to grow up too soon, or to be thrusted out of school quickly with a degree, but without a passion or practical means to begin "the next life". I may not be graduating in time, but I am graduating with a discovered passion and a sense of purpose. For this I want to thank The University of Helsinki and in particular professors Klaus Kultti and Markku Lanne for support (and often their flexibility). Special thanks to Will Phan for an illuminating Advanced Micro course and for his help in understanding choice problems.

A belated graduation is the cost of a job started before this thesis. For their confidence, support, flexibility and understanding I want to thank Katri Kennedy, Sanna Viipuri and Izabela Hollingworth, without whom this thesis would still be an idea on a PowerPoint slide.

For their help with statistics and econometric modeling, the biggest thanks to Tor Martin Christensen and Rainer Avikainen. Special thanks for help, not only with the econometric section of this thesis, but with most of my Master's degree (and all Weird Things in general) go to the peculiar Jirka Poropudas, Ph.D. Thanks to the online R-community and YouTube – lecturers.

Thanks to the game of basketball. Not only is it my true passion in life but it has gifted me with experiences and friends to cherish forever. Thanks to my teammates and coaches, without whom I would be miserable.

Finally I want to thank my dad Pasi, my sister Lilli and my mom / best friend, Kaisa. You have encouraged me to do what I love and granted me the freedom to make decisions – often stupid ones – while never showing anything but love and support. For that I am truly grateful. I owe all to you.

CONTENTS

List of Figures

1 Introduction

In this thesis, customer churn is analyzed among movers in the Finnish insurance market. Econometric modeling is conducted based on a sample of an insurer's customer data to find variables (i.e. characteristics and preferences) that predict churn among customers around the time of their move. The theory upon which this modeling is based was pioneered by Nobel-laureate Daniel McFadden and his contemporaries in the field of discrete choice.

Per Posti (2017a), close to one in five Finns moved homes in 2016. Moving homes, i.e. changing from one address to another triggers the moving individual to proactively take action on two activities:

1. File an official change of address with the Finnish Postal service (Posti).
2. Renew all address-based subscription services, such as electricity, internet and home insurance. This thesis focuses solely on insurers and their customers.

Activity 1 enables Posti to be a reliable authority on address changes in Finland. Posti keeps records, but also does business by providing data and marketing opportunities to providers of the subscription services described above.
Activity 2 leads to moving being a crucial moment in any continuous customership, as the product related to the old address will be terminated, giving the moving customer the opening to explore options elsewhere.

When a customer cancels their existing subscription in favor of a competitor, it is said that customer has churned. Minimizing churn is as important to an insurer's business as sales, with the distinction that retaining customers is more cost effective than acquiring new ones. Therefore understanding who the likely churners are - based on available characteristics and revealed preferences - allows the insurer to accurately and appropriately target high churn-risk individuals out of a large mass of movers, to minimize churn and subsequently improve result.

The work begins with a review of basic concepts of risk and insurance in Chapter 2. Hillson & Murray Webster (2007) provide a compact breakdown of risk and management thereof, while Vaughan's *Fundamentals of Risk and Insurance* (1996) complements by defining key concepts in the insurance business along with a look back into history at what preceded the modern industry. As this thesis focuses on moving and insuring specifically in Finland, it is important

to present the modern insurance market in Finland in greater detail. The annual report on the insurance industry, published by Finance Finland (2017) serves as foundation for the introduction of key indicators of insurance business as well as the competitive situation in the market as of 2016.

Among the key performance indicators reviewed in Chapter 2 is retention and its counterpart – churn. Chapter 3 details the event of moving homes as a distinct driver of churn, not only in insurance but all fields of business offering address-related services on subscriptions (electricity, internet, etc). Publications by Posti (2016, 2017a, 2017b) serve as the factual basis for this chapter. Combining the model of insurance business with what we know about moving customers allows us to formulate the research problem with guidance from Jaccard & Jacoby (2010): what kind of customer characteristics and preferences indicate an increased risk of churning? Understanding and identifying high-churn-risk individuals allows the insurer to be more efficient in retention-improving customer save -activities.

Before creating a model and experimenting, the scientific basis must be established. In Chapter 4 previous work in the field of choice theory and modeling is reviewed. Churning is a choice between two alternatives; an address change forces the customer to explicitly choose whether to remain with their current insurer or churn and join a competitor. Papers detailing the application of choice theory into churn problems in the fields of telecommunications and internet services are presented (Madden et al. 1999, Kim 2004, etc.), but the theoretical backbone of churn models – as the one created in this thesis – is the comprehensive work of Nobel laureate Daniel McFadden. Chapter 4 presents the early, pre-McFadden era choice theory (Pareto 1906, Marschak & Block 1960, Hartley 1996) as well as McFadden's own research (1974, 2000) that together with the improved availability of individual data, shaped modern choice modeling. In essence, the modern binary choice model – such as a churn model - uses available individual characteristics and preferences towards service attributes to assign a probability of churn for each individual. The regression yields weights and significances for each variable (characteristics and preferences) that can help segment the mass of customers based on their churn probabilities.

The churn modeling is conducted in Chapter 5. I offer a sample of 50 000 individuals, provided by a Finnish insurer for discussion. The data is manipulated by individual analysis of variables, based on which 21 appropriate descriptors remain. Additionally the sample is subsetted to only

include customers who had an active home insurance before their address change. The remaining sample consists of 24 230 observations.

In order to capture the full volume of churners it is necessary to run three experiments.

1. All customers who canceled their home insurance are treated as churners. The model is created by regressing home churners on specified variables.
2. The second model is created by treating customers who canceled all insurance products as churners. Full churn is regressed on the specified variables out of a data of home insurance churners.
3. Finally, the third model regresses full churn on the specified variables with the full sample of 24 230 movers.

The experiments revealed variables indicating economic status to be important determinants of churn in all three experiments. Customer age and customership tenure (duration) were significant and considerable predictors, with older and more long-term customers being less likely to churn. In addition the presence of optional personal insurance as well as contact with one's insurer prior to the address change significantly increased the probability of retention.

The results of the experiments were significant and deemed to fit their respective data. The predictive power of the models (with the exception of Model 2) was quite small. This leads for the research to conclude that there exists some variable or variables affecting churn probabilities more than what is captured by the data used here. As the data does not account for prices, price changes or competitor activities it is reasonable to assume they would affect churn probabilities significantly. In Section 5.3 and finally in Chapter 6, conclusions and discussions are offered with critique of the work.

2. Insurance: History and Industry Overview

More things might happen, than will happen.

Plato (427-347 BC)

2.1. Uncertainty, risk and insurance

When faced with uncertainty we as humans do not always act rationally (Hillson & Murray-Webster 2007). The need for insurance is born from an individual's aversion to risk. By definition, insurance is a contract with which the insured assumes a certain, smaller loss in exchange for a guarantee to be compensated for a greater, uncertain loss. This relation, as described by Vaughan (1996), can be written as follows:

Buying insurance is choosing to lose a little at probability 1 over losing a lot at probability $p < 1$, i.e. the chance of the agreed upon risk being realized. The insurance business is based on the irrational nature of risk aversion, compounded by the asymmetric nature of information regarding risks. Assume an individual owns a house and wants to insure that house for in case it is destroyed by fire sometime in the future. The house is worth *m* amount of money. The guaranteed payment – the premium - made by the individual is set by the insurer and is denoted by *y*. In perfect competition, the premium is set at

$$y = p * m$$

Over one contract period, the insurer either gains *y* or loses *y-m*, depending on whether p_0 is realized. As the contract period, i.e. a year, is renewed, the central limit theorem states the insurer's profit would converge to zero.

The irrationality Hillson & Murray-Webster discuss lends credibility to the idea that even in a setting of perfect and symmetric information between insurer and insured about risk p_0, the insured would agree to a higher premium. The information, however, is not symmetric in practice. The correct value for *p* is impossible to know for certain in advance. This allows the insured and the insurer to have very different predictions of what *p* is. As the insurer has gained historical data from previous contracts on the actual realizations of *p*, it is reasonable to assume they have a more informed and accurate prediction of *p*. The premium is now being set as follows.

$$y = p_1 * m > p * m$$

Here, p_1 represents the insurer's assessment of p and p_2 the same for the insured. As the insurer has better information than the insured, they know to set y at the highest possible p_1 the insured will still accept, so long as $p_1 \geq p_0$. Over one contract period, the insurer's profit is again contingent on the stochastic nature of the risk, but over many contract periods, this premium would yield positive returns. Thus both irrationality when facing uncertainty and asymmetric information allow for insurance to be a profitable form of business for the insurer. Professor Howard Kunreuther and his co-authors describe the bi-polar nature of uncertainty in their 2013 publication *Insurance and Behavioral Economics* (Kunreuther 2013). As is evident from the previous example, when the premium is set properly, it is crucial for the insurer to hold on to their customers for many contract periods in order to minimize noise.

2.2. The insurance business model

In traditional retail, the supplier knows what each unit of good has cost to produce and its goal is to sell each unit for a price that, at minimum, covers that cost (e.g. Varian 1987). If the only costs supplier A, a baker, has in production are two euros per bread and he can sell each bread for three euros, his profit earned grows with each bread sold.

$$\pi = pq - cq$$

As c, the unit cost, is known to the baker, it is in their interest to sell as much as possible, as long as the price p in the market is greater than c.

$$q, \pi \ \rightarrow \infty, when$$

$$p > c$$

An insurer's revenue is the premium paid by those insured. The insurance contract in itself is immaterial, it costs nothing to produce. The cost for the insurer arises, when an insured suffers a loss covered by the insurance contract. The request by the insured to be compensated is called a *claim filed* and the money moving from insurer to insured is called a *claim paid*. With these definitions, the insurer's model of business in the form of the baker's profit maximization equation, is

$$\textit{Profit } (\pi) = \textit{premium earned} - (\textit{claims paid} + \textit{overhead})^1$$

[1] The overhead for an insurer include e.g. employee compensation, rental expenses, etc. In this paper, overhead will be treated as a constant positive, and denoted by *OC*.

Insurers strive for profitability. When observing an insurance contract – a policy – over time, Vaughan (1996) defines that policy as having been profitable, if the above equation is positive. If the costs, on the other hand, exceeded the premium earned, the policy was unprofitable. In the example above, with the baker selling his bread, there is no uncertainty to the cost-side of the equation. The baker's only concern is how many units he manages to sell. This is quite different from the insurer, as he cannot know in advance what the costs arising from a policy will be. Uncertainty is therefore not only the cause of the need for insurance, but also the thing that most separates the insurance business from traditional retail.

Upon entering an insurance contract, the insurer does not know

1. Whether the insured will suffer a claimable loss, or
2. How great that loss would be.

The price for the contract, the premium, is still agreed upon at the signing of the contract, so the best the insurer can do is use statistics and probability to forecast both 1 and 2 and set the premium accordingly, to remain profitable over time. Contrary to the baker, an increase in sold units does not always guarantee a better result. In the case of an unprofitable policy, it is quite the opposite – when a policy is priced too low relative to the risk, the insurer's negative profit grows greater with each policy sold over time. An insurer's performance cannot therefore be measured purely by revenue, as it says nothing of the costs. Instead, insurers are assessed by ratios that capture the cost-side of their business as well. Some key indicators are presented below, as defined by Vaughan (1996) and Finance Finland (2017).

- **Loss Ratio (LR, sometimes Claims ratio)** divides the claims paid out during a period's (often a year) time by the premium earned. *LR = Claims paid / Premium earned*
- **Cost Ratio (C, sometimes Expense ratio)** captures the overhead (OC), as the premium earned must cover those costs as well. *C = OC / Premium earned*
- **Combined Ratio (CR)** is the key indicator of an insurer's performance. It combines both the loss ratio and the cost ratio and when it is below 1, business is profitable. *CR = LR + C.*

Combined ratio yields an insurer's result, sometimes referred to as the technical result. For reference, Figure 2.2.1 presents the Finnish P&C –market development according to these measures.

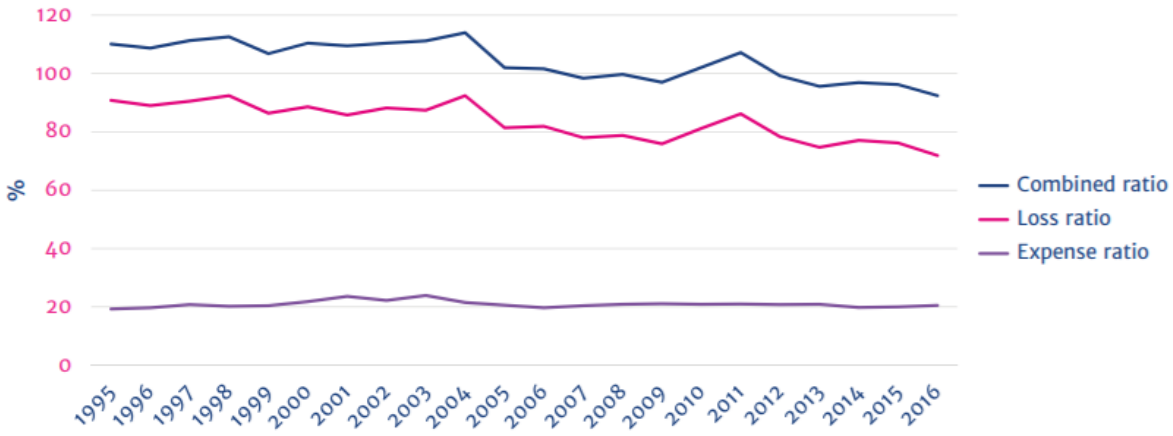


Figure 2.2. 1 The development of the Finnish P&C Insurance market.

In 2016, the Finnish P&C market as a whole reported 4,25 billion euros of premium earned. The corresponding volume of claims paid was 3,06 billion euros. Overhead expenses were 0,87 billion euros. These values add up to a combined ratio of 92,41 %, which in turn yields a positive result of 323 million euros for the market. Please refer to Figure 2.2.2 below.

Premium earned	Claims paid	Expenses	Loss ratio	Cost ratio	Combined ratio	Result
4255	3061	871	71.94 %	20.47 %	92.41 %	323

Figure 2.2. 2 The Finnish P&C Insurance market key figures in 2016 (MEUR)

An insurer, as any rational agent on the supply side of a market, maximizes their profit, i.e. their result. Result maximization is a delicate balance between increasing premium earned, without compromising the combined ratio.

2.3. The insurance market from Antique times to modern Finland

2.3.1. A brief history of insurance

Vaughan (1996) traces the history of insurance back millennia, all the way to Antiquity. Chinese merchants would share risk by distributing the goods they were shipping on each other's vessels. If one merchant's boat would crash in the rapids, the loss would be spread out and shared by all, instead of being carried only by one unlucky individual. Although risk-sharing techniques akin to the above example existed in the ancient world, modern insurance

business began to take shape only after the commercial revolution in Europe, following the crusades.

Insurance is generally divided into two categories: personal (life and casualty) insurance and property-and-liability insurance. Personal insurance protects against financial loss in matters of life and health, whereas property and liability insurance safeguards individuals from costs arising from damages or peril to material possessions or assets. Modern property insurance has its roots in marine and fire insurance, originating in the thirteenth and sixteenth centuries, respectively. Casualty insurance as we know it began to take shape centuries later, in the 1800s.

Finland today categorizes insurance slightly differently, the main reason for this distinction is written in the Insurance Companies Act (2008). From distinguishing between personal and property, Finland divides insurance into *life* and *non-life* products. Life insurance is used to financially compensate beneficiaries of the insured in case the insured passes away. Non-life, also known as property and casualty (P&C) insurance, on the other hand consists of insurance covering not only property and liability, but also financial loss from health-related matters, death not-withstanding.

2.3.2. The Finnish insurance market

At the end of 2016, there were 52 licensed insurers operating in Finland, out of which 36 were *property and casualty* –specialized (P&C) companies, per Finance Finland (2017). Despite the relatively large number of suppliers, 99 % of the non-life premiums in the market in 2016 were written by only eight companies. Moreover, 81 % of the market was controlled by the three largest insurers: OP, LocalTapiola and If P&C.

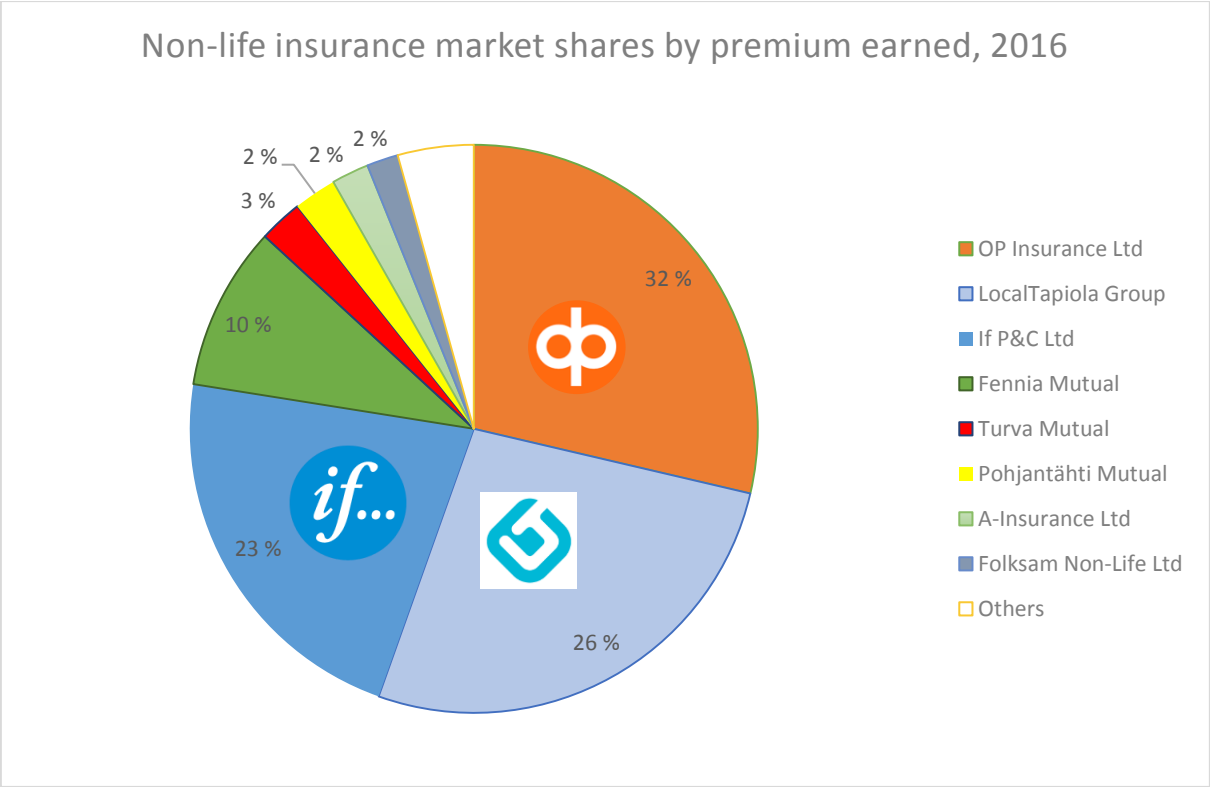


Figure 2.2. 3 The market shares of Finnish P&C Insurers, 2016

When a customer purchases an insurance policy to protect their assets or their self from financial loss, it is equivalent to saying the insurer and the insured agree upon and enter a contract. In Finland, by law, when a contract is entered, it is valid continuously until either party actively terminates it. At the point of purchase, the insurer and the insured agree to an annual premium, which is then charged every year, unless the insured decides to terminate the contract. The insurer may also terminate the contract for certain reasons, such as failure to pay the premium or based on underwriting-guidelines regarding the insured individual's age. It is important to note that the insured may terminate the contract at any point in time, but the insurer may do the same only once per year, on what is called the date of renewal.

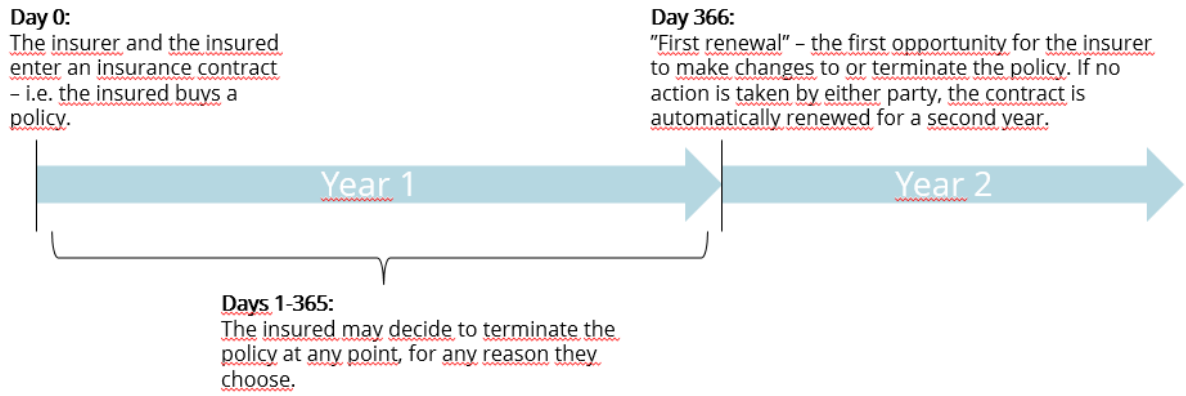


Figure 2.2. 4 The Finnish insurance contract

2.3.3. Retention

As previously discussed, the insurer maximizes result by managing both the premium earned and the combined ratio – in other words revenue and profitability. Revenue is generated from both individuals purchasing new policies but also existing policies being renewed at renewals. It is important to once again point out the importance of uncertainty when assessing profitability. As the price for the policy is agreed upon beforehand, the insurer relies on past experience, recorded data on not only the insurable asset or person (known as the exposure), but on the customer (insured, policyholder) as well. It is reasonable to say that assessing the risks related to a customer is made easier in the case of renewal, as the insurer already has data on the customer and the exposure at hand. A renewal also happens automatically, so renewing a policy for another year costs nothing to the insurer. Renewing policies – called retention – is therefore less risky and requires less effort on the insurer's part, then creating revenue through new customers or policies. Many publications on retention (e.g Ahn et al 2006) have explicitly stated that retaining existing customers is cheaper than acquiring new ones. It is therefore crucially important for an insurance company to ensure profitably priced customers are retained. I will define some measures important to insurers, based on the nature of the Finnish insurance contract as well as the Insurance Contracts Act (1994).

- **Retention rate**: Used to measure retaining of customers. As a single individual may have multiple policies purchased from the insurer, rates defined this way fails to capture those customers who only terminate one/some, but not all of their policies. Retention rate is always observed over a fixed-time and presented in a form akin to the examples listed here:

- *12-month retention rate:* A random anchor-date from the past is selected. All customers who have at least one valid insurance policy at that date are flagged. These flagged customers are observed exactly 12 months later, and those who still have at least one valid insurance policy are the ones who have been retained. The retained customers are then divided by the original flagged customers to generate the 12-month retention rate.

$$12m\ retention\ rate = \frac{Nr\ of flagged\ customers\ observed\ 12\ months\ after\ anchor\ date}{Nr\ of\ customers\ flagged\ at\ anchor\ date}$$

- *First year retention rate*: First year retention is calculated by looking back exactly one year. Out of all new customers who bought their first policy one year ago to the day, the ones who still have at least one policy valid are the retained ones. The number of retained customers is divided by all the new customers entering a contract exactly one year ago.

$$1st\ yr\ retention = \frac{Nr\ of retained\ from\ 12m\ ago}{Nr\ of\ new\ customers\ 12m\ ago}$$

2.4 Churn

As the law states, a customer is free to terminate an insurance policy at any time, which means that every customer poses a flight risk at all times. When such risk is realized and the customer terminates their contract with the insurer, it is said that the customer has churned. The traditional definition of churn in a subscription-based service model is as simple as it is described above. If a customer continues their subscription at some observation period, the firm has retained that customer. On the other hand, if another customer no longer subscribes at that same observation period, the customer is said to have churned. It is important to note that in the Finnish insurance market, customers often have more than one policy purchased from an insurer. This leads us to create a distinction between *partial churn* and *complete churn*.

If a customer terminates <u>all</u> of their existing contracts, leaving no policy subscriptions active, we say that customer has *churned completely (or fully)*.

> If a customer terminates at least one of their existing contracts, leaving at least one policy subscription active, we say that customer has *partially churned.*

Retention analysis should account for both for the following reason: focusing only on complete churn will not capture the entire decrease in revenue. Consider an extreme example where all customers have two subscriptions with an insurer and all policies cost the same. Now consider all customers cancel one of their two policies, leaving one active. If we were to measure retention simply with the definition of complete churn, the firm's retention would appear to be 100 %, yet the revenue received from premiums would drop to half from the previous period. For this reason, the models constructed in this thesis will account for both partial churn and full churn.

Improving retention is synonymous with decreasing churn. When each customer is free to churn at any point in time, it is useful to look for spikes in churn rates. What kind of events precede churn? What connects churning customers? To be able to take action in improving retention, one must know *who* the customers are that decide to leave and *why* they decide to do so. Following these thoughts, Chapter 3 presents the research problem. Chapter 4 presents theory and Chapter 5 the empiric models to solve said problem.

3 Research problem: Churn among movers

We now know that managing churn is as, if not more, important to result maximization as new customer acquisition. As churning can happen literally at any point during a year, it is reasonable to assume retention rates remain quite stable regardless of the anchor date. A valid question to ask is whether there exists a common event – some experience or activity around which churn rates have increased. This chapter presents the move event (moving from one address to another) as a Finnish phenomenon, around which churn-rates in subscription-based services, such as insurance, tend to spike. A study conducted by the Finnish Postal service (Posti 2016) is presented, and the research question is formed based on that study. This thesis asks *what are the characteristics and experiences for moving customers that indicate an increased probability to churn*.

3.1. The move event

According to Posti (2016), 700 000 households move homes each year, with those households consisting of approximately one million individuals. This represents roughly 18 % of the Finnish population. Moving is reported and recorded as a change of home address to the Finnish Postal Service (henceforth Posti). Since Posti is the official authority receiving the reports of moving, it is a reliable source of movers-related data.

Close to one in every five Finns moving within a year seems like a high volume. Why do Finns move so often? Posti (2017b) offers the economic environment to be a reason. In years of recession 2008 and 2009, when the GDP growth rate shrank from 5,2 % to -8,3 %, around 650 000 households reported an address change. Years 2015 and 2016 however, when GDP growth rate increased from -0,6 % to 1,9 %, saw annual moving numbers of 710 000 and 720 000 respectively – a growth of nearly 11 %. [2] The high volume of movers is driven mainly by individuals under the age of 30 (53 %) and 61 % of movers moved into apartment buildings. (Posti 2017b).

[2] GDP figures from Official Statistics of Finland (OSF) (2016). Annual national accounts [e-publication]. ISSN=1798-0623. 2016. Helsinki: Statistics Finland [referred: 9.12.2017].
Access method: http://www.stat.fi/til/vtp/2016/vtp_2016_2017-07-13_tie_001_en.html

Posti (2016, 2017a) offers the data most relevant to subscription-providers, such as insurers. An address change means that one must actively update their address-tied subscriptions, such as electricity, internet and home insurance. The discontinuation in the subscription, caused by the move creates an increased risk of churn. This is intuitive. A property of continuous services such as insurance or internet, is that a consumer does not actively consider them on a daily basis. When faced with an event where it is necessary to re-visit your subscriptions and one is again presented with the prices of those services, it is reasonable to consider competitors' offerings as well. "Tendering your subscriptions" has become customary whenever an address change is imminent. Posti knows this and does business with the insights they have on movers. As was touched on in Chapter 1, Posti provides marketing visibility and *leads* on moving customers for businesses looking to either protect their own moving customers and/or win movers over from competitors. When testing the online change-of-address service on Posti's website, it is evident most if not all electricity and insurance providers have taken advantage of these services. It is reasonable to assume this is another factor in the high churn rates around the moving date.

This intuition is backed by data. Data provided by Finnish Insurer X shows that when the moving date is set as the anchor date, 30-day churn rates are multiplied, than when measured on a random arbitrary anchor date. To the insurer, this means that (assuming the mass of movers is distributed among companies according to market share) 18 % of customers present a common, drastically increased risk of churning at some point during a year.

3.2. Who churn and why?

The insurance customers with an increased risk of churning, the movers, have a defining characteristic in common: they change addresses while moving houses. However most things are very different from mover to mover, not least of which the outcome of the move in regard to retention. Some choose to continue their subscription with their original insurer, while some choose to churn and select another insurer. Why? Analyzing customer data on characteristics and experiences should provide additional insight into what connects churners to one another and what separates them from non-churners. This thesis asks whether customer data can be used to accurately predict those insurance policyholders who, when triggered by an address change, would choose to churn and choose another insurer. If the answer is "yes", then naturally it is of interest to find the characteristics and events in a customer's profile that significantly

either increase or decrease the probability to churn. The following chapter presents the theoretical basis for answering that question, detailing past work on modeling choice between discrete alternatives. Chapter 5 details the econometric process and its results.

4 Theory review: Discrete choice modeling

Discrete choice modeling is a relatively young field of economic study. It was developed into its modern form in the latter half of the 20th century by econometricians trying to answer questions in travel and transportation analysis. The foundation of their work, however, can be traced back to some of the more fundamental results in microeconomics, such as the classical theory of the consumer. In this chapter I present some of the notable research done on discrete choice. Along with the different approaches and models applied to economic problems today, I will begin by reviewing some of the groundwork upon which economists such as Daniel McFadden and Kenneth Train began to model and predict individual consumer choice among discrete alternatives. Finally I present some contemporary applications of choice theory into churn problems, most notably Madden et al's (1999) study of subscriber churn in the Australian ISP market.

4.1 From utility to modern churn models

In 1985 Moshe Ben Akiva and Steven Lerman described discrete choice analysis as modeling an individual's choice from a set of mutually exclusive and collectively exhaustive alternatives (Ben Akiva & Lerman 1985). A continuous choice problem asks *how much* you will choose from a continuous consumption set, an example of which is the open set (0,1). A discrete choice problem rather poses the question of "which one?" or "how many?" will you choose from, for example, the set {1, 2, 3}.

Understanding and rationalizing decisions made by an individual facing discrete alternatives has been the subject of study by economists and psychologists alike. A natural starting point in studying choice is to ask *why* one thing is chosen over something else. It is a cornerstone of microeconomics that a sensible individual will choose the alternative that yields him more satisfaction, or utility, than any other available alternative. This concept of a consumer selecting the alternative with the highest resulting utility can be found already in some of the very first works on mathematical economics, such as William Jevons' opus, *The Theory of Political Economy* (Jevons 1871). With mathematical tools being introduced, it became relevant to measure and give values to utility levels. Building on the work by Jevons, Vilfredo Pareto

proposed in 1906 the concept of ordinal utility. Ordinal utility states that merely the rankings of utilities matter, not the actual values. Consider an individual making a choice between two alternatives *a* and *b*. If the utility resulting from consuming *a* is greater than that of consuming *b*, i.e.

$$u(a) > u(b)$$

then we say the individual *prefers a* to $b \Leftrightarrow a \succ b$. Consider now that we know

$$u(a) = 8, \qquad u(b) = 4$$

Ordinal utility states that knowing the respective values of utilities gives us no more insight to the individual's preferences, than what we already had. We can still only state that $a \succ b$, but nothing of the intensities of preference (Pareto 1906).

With the framework established, economists began to examine ways of using the utility maximizing consumer to model market-level behavior. The most notable way of aggregating consumer choice was to use an individual's demand to represent the population. James E. Hartley presents and critiques this method, called the representative agent, as having roots in Alfred Marshall's representative firm theory first introduced in 1890. All variations from the representative agent in observations among individuals were represented by an additive disturbance, i.e. an error term. The variation was attributed rather to measurement errors than unobserved factors in individual agents. Hartley argues that it is extremely hard to give a consistent and empirically usable definition of what a representative individual is like, stating that no representative agent can model heterogeneity (Hartley, 1996). The representative agent model is presented below.

4.1.1 The representative agent model (McFadden 2000)

Consider a consumer whose preferences are represented by $u(\boldsymbol{x})$, where $\boldsymbol{x}$ is a vector of consumption levels of various goods. The consumer maximizes this utility subject to a budget constraint $\boldsymbol{px} \leq w$, where $\boldsymbol{p}$ is a vector of prices and w is the consumer's income.

The consumer's demand function is then:

$$\boldsymbol{x} = d(w, \boldsymbol{p})$$

The market level demand is

$$\boldsymbol{x} = d(w, \boldsymbol{p}) + \varepsilon$$

where ε is the disturbance added to account for variation among observations.

As critique of the representative agent grew, focus turned to the heterogeneity of preferences. Preferences vary, because individuals value different attributes in commodities. This understanding, along with a rapidly growing availability of individual consumer data paved way for more accurate modeling of individual choice. Econometrician Daniel McFadden, whose work serves as the theoretical basis for the upcoming experiment, won the Nobel Prize in 2000 for his pioneering work in modeling discrete choice. In his prize lecture, he summarized the idea upon which choice analysis is based as follows:

> "The heart of the standard or rational model of economics is the idea that consumers seek to maximize innate, stable preferences whose domain is the vector of quantities and attributes of the commodities they consume." (McFadden 2000)

McFadden approached discrete choice from a transport economist's perspective. In his 1981 paper *Econometric Models of Probabilistic Choice* he presents an example of binomial discrete choice: the choice of commuting to work by car or by bus. Upon this example, Madden, Savage and Coble-Neal (1999) applied discrete choice theory to a churn probability model in the Australian internet service provider (ISP) market. As established in chapter 2, internet services are similar to modern Finnish insurance policies in the sense that both are continuous subscriptions that can be terminated by the customer at any moment. The paper by Madden et al therefore functions as an appropriate, scientific basis for creating a churn model for an insurer. Ahn et al (2006) and Kim (2004) constructed churn models in a similar way to McFadden in the Korean mobile communications industry.

Whether to churn is a choice an insured customer makes between two discrete alternatives – staying or churning. The choice, as presented by Madden et al (1999), originally proposed by McFadden (1974), depends on the customer's characteristics and their valuation of the insurer's attributes. Policies are continuous, as explained in Chapter 2, so every moment of every day the subscription continues, is in fact the customer making a choice to stay. This is rather abstract, as it is quite obvious individuals do not make a conscious decision to stay with their insurer each moment of each day. We are, however, studying a sample of customers with a very distinct event in their customership – the event of moving homes. Because an address change effectively terminates the home insurance policy linked to the old address, the moving customer is in fact faced with a decision between two alternatives – to continue with my current insurer at my new address, or churn and start a policy with some other insurer. Following Madden et al. closely, we can formally state the customer's problem. The decision can be modeled by defining the n^{th} customer's satisfaction with their existing insurer, *insurer j*, with the customer's available characteristics (s_n) and *j*'s measured attributes ($z_{j,n}$). The indirect utility function takes the form:

$$U_{j,n} = U(z_{j,n}, s_n), \quad j = \{churn, stay\}$$

A customer's probability to churn is equivalent to the probability that selecting a new insurer while changing the address would yield them higher utility than staying. Or:

$$P(churn|j) = P(U_{churn,n} > U_{stay,n})$$

The model by Madden et al, relates a binary variable (churn = 1, stay = 0) to the measured characteristics and attributes in a logistic regression (as detailed in Figure 4.1.2):

$$P_{churn,n} = \frac{e^{\beta x'_{j,n}}}{1 + e^{\beta x'_{j,n}}}$$

$$P_{stay,n} = \frac{1}{1 + e^{\beta x'_{j,n}}}$$

Where $P_{j,n}$ is the probability of churn for the n^{th} customer and $x'_{j,n}$ is the vector of measured *z* and *s* for *j* and *n*, respectively. β represents the parameters to be estimated. This model yields weights and significances for the attributes and characteristics used as regressors in the model. (Cox 1958). Barring statistically insignificant results or a poor goodness of fit of the model,

results from an experiment such as this can be used to identify and segment moving customers based on churn risk, to prioritize customer saving activities accordingly.

Modern discrete choice modeling can therefore be interpreted as using individual data on characteristics and preferences to give individuals a probability of choosing something.

Figure 4.1.2 The binomial logistic regression model, proposed by Cox (1958) and applied by McFadden (1974)

Consider Y_i to be a binary process. Then $Y_1,..,Y_n$ are random variables, each taking either the value "1" or "0".

Let $x_1...,x_n$ be a set of fixed numbers.

In the simple form of the binomial problem, we suspect x_i to have a relation with the probability of $Y_i = 1$, denoted here as μ_i. The linear relation between μ_i and x_i would be

$$P(Y_i = 1) = \mu_i = \alpha + \beta x_i$$

A linear relation would produce unusable estimates, as μ_i is a probability and thus must fall between 0 and 1. By the logistic law we have

$$logit\ \mu_i = \log\left\{\frac{\mu_i}{(1-\mu_i)}\right\} = \alpha + \beta x_i$$

Further,

$$\mu_i = P(Y_i = 1) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

And

$$(1-\mu_i) = P(Y_i = 0) = \frac{1}{1 + e^{\alpha+\beta x_i}}$$

Estimates of β_i can therefore be interpreted as the proportional change in the *log-odds* of "success" in the binary process. For the proportional change in the true probability, the operation $exp(\hat{\beta})$ must be performed.

5 Empirical application: Modeling churn in the Finnish insurance market

This chapter presents the application of discrete choice theory into practice to answer the research problem presented in Chapter 3, i.e. *which customer characteristics and preferences imply an increased risk to churn at the time of moving homes.* A sample of customer data (provided by Finnish Insurer X) is used in the experiments. Section 5.1. presents and discusses the data in detail, along with justifications for data manipulation and cleaning. In 5.2., the binomial logistic regression model is constructed as described in the previous chapter and its results presented. Altogether three experiments are conducted to find significances for:

1. Characteristics and preferences indicating an increased probability to change their home insurance.
2. Characteristics and preferences among those who change their home insurance, indicating complete churn.
3. Characteristics and preferences indicating complete churn in the mass of moving customers.

Finally, 5.3. presents model evaluation along with discussion of the results.

5.1. The Data

The data used in this thesis was kindly provided for modeling by a Finnish insurer, henceforth known as Insurer X. It is a sample of the insurer's customer data, consisting of information the insurer registers and stores about each individual and their customer history over time. Important note: the test sample has been cleaned of all identifiable personal and company information. The reason for this being the protection of the customers' data privacy as well as the insurer's private, business-sensitive data.

The sample was created as follows: all such customers who filed a change of address during the year 2016 were selected in the sample.[3] This sampling method ensures that the individuals selected are in fact those who were customers before moving, but in no way discriminates based on whether a customer chose to remain or churn after the move. The 90 day –window is

[3] Definition of customer: "an individual with a valid contract for at least one valid insurance policy 90 days before the date of reported change of address"

arbitrary and is in place to ensure as much of the effect of moving on churn is captured. The columns include variables available to the insurer, depicting customer characteristics and preferences, with the aim of capturing similar characteristics and preferences Madden et al (1999) obtained by survey.

The sampling yielded 50 000 observations. Each observation – or row – represents an individual customer. The variables – or columns – are customer characteristics or historical nodes of information having been accrued over the length of each individual's time as a customer. There are altogether 21 variables. They range from demographics, such as age group to very detailed information related to experiences as a customer, such as "how many visits has the customer made to the online service in the last 45 days before sampling?". Table 5.1.1 presents the variables in categories, as Madden et al (1999) did, along with explanations.

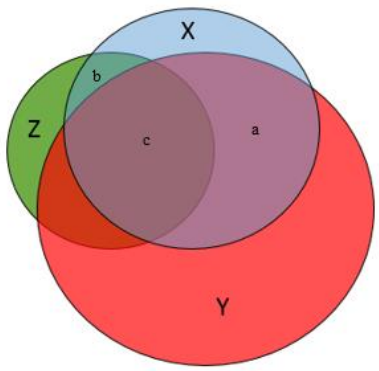
As established in Chapter 3, moving homes leads to an individual having to take action on their home insurance. Therefore our data should only include such customers who had an active home insurance 90 days prior to the registered change of address. Our original data, *N*, contained 50 000 observations. Figure 5.1.1 presents data set N as a union of three subsets – *X*, *Y* and *Z*, where each subset contains observations with at least one active insurance policy in the three lines of business - property, motor and personal – respectively.

$$N = X \cup Y \cup Z$$

With this definition, our dataset of interest is the subset $O \subset N$, so that

$$O = X \cup (X \cap Y) \cup (X \cap Z) \cup (X \cap Y \cap Z).$$

a b c



5.1. 1 The subsetting of data from N to O

The subset *O* consists of 24 230 observations.

ECONOMIC (preference)		
payment_plan	*categorical*	Customers choice of payment installments: 1, 2, 3, 4, 6
nr_lob_motor_90_days_ago	*categorical*	Number of motor objects: "0", "1-2", "3-4" and "5+"
nr_lob_property_90_days_ago	*categorical*	Number of property objects: "SINGLE" and "MULTIPLE"
nr_lob_personal_90_days_ago	*categorical*	Number of personal objects: "0", "1", "2" and "3+"
SOCIO-DEMOGRAPHIC (characteristics)		
age_group	*categorical*	Age of customer: "under27", "28-38", "39-50", "51-64", "65+"
Gndr_Cd	*categorical*	Gender of customer. Binary "MALE" / "FEMALE"
USAGE (preference)		
Web_visits_last_45_days	*numeric*	Customers visits to insurer's digital platforms.
Ctc_email	*categorical*	Has customer given consent to receive e-mail? "YES"/"NO"
Ctc_phone	*categorical*	Has customer given consent to receive call? "YES"/"NO"
number_of_inbound_calls_90d	*numeric*	How many x the customer has called the insurer over 90 days
Ins_Polcy_Dlv_Cd	*categorical*	Document delivery method: digital or paper.
duration	*categorical*	Customership length / tenure. "1 yr", "2 yrs", "3-5", "5-10", "10+"
EXPERIENCE (preference)		
tm_contacted	*categorical*	Has the customer received a call in the last 90d? "YES" / "NO"
Customer_Commented_Eff	*categorical*	Has the customer commented on a CX-survey for effort "YES" / "NO"
Customer_Commented_Imp	*categorical*	Same as above, but for improvement "YES" / "NO"
CX_forms_answered_last_90d	*numeric*	How many CX-forms the customer has answered to the last 90 days.
CX_forms_received_last_90d	*numeric*	How many CX-forms the customer has received the last 90 days.
DMs_last_90d	*numeric*	How many pieces of marketing mail the customer has received the last 90 days
eDMs_last_90d	*numeric*	How many pieces of marketing e-mail the customer has received 90d.

5.1. 2 The 19 regressors to be used in initial Experiments 1-3.

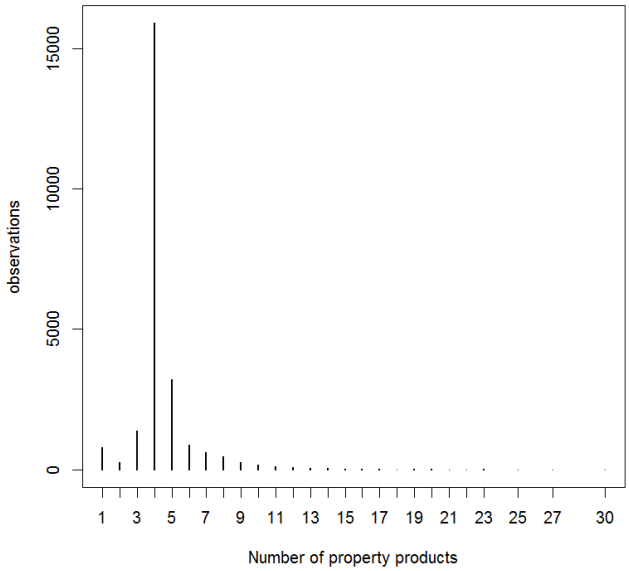
5.1.1 Notes on variable manipulation and combining

The initial data set combined a larger number of variables than what is detailed in Table 5.1.1. Some were disregarded right away because of missing values accounting for more than 90 % of the observations (with no significance on churn in the remaining observations) and some were bundled to create new categorical variables. This section details the key manipulations that were conducted in order to create more accurate models, the R-code to which is found in the Appendix.

The variable *payment_plan* is a categorical variable with five levels – "1", "2,", "3", "4" and "6" – each indicating the customer's choice of payment installments. It is created by combining five separate numeric variables denoting the number of bills with said installment option the customer has.

Duration is a categorical variable denoting the length of the customer's tenure at their current provider. It has five levels "1 yr", "2 yrs", "3-5 yrs", "5-10 yrs" and "10+ yrs". The original variable was a numeric variable indicating how many days a customership has been active.

The variables indicating the size of a customer's insurance portfolio - "*nr_lob_[lob here]_90_days_ago"* – were originally numeric variables, the values of which indicated the number of active policies a customer had in that line of business. Using these numeric variables as regressors yielded absurd results caused by, as it turns out, Insurer X:s way of accounting for policies. For example the variable indicating the amount of property products (i.e. home insurance) the customer had, contained by far the most observations at value 4, with only a fraction at 1-3 (Figure 5.1.1). This is because a standard home insurance contains four covers. In our data, four "objects" make one product. Due to this, the variable was converted to categorical type, with observations ≤ 4 being registered at level "SINGLE", while all those > 4 registered as "MULTIPLE". A similar conversion was done for LOBs personal and motor.



5.1. 3 The distribution of observations by property products in the original variable

5.2. The Models

After manipulation and cleaning, the sample now consists of xxxx observations with xx variables. Per Madden et al (1999), a binomial probit model is used to relate the probability of a customer leaving their insurer with variables depicting economic and demographic customer characteristics as well as those depicting customer experience and preferences (usage). The full list of variables along with explanations is provided in Table xx. From Chapter 4, the model is written as:

$$P_{churn,n} = \frac{e^{\beta x'_{j,n}}}{1 + e^{\beta x'_{j,n}}}$$

Chapter 3 detailed the distinction between partial and complete churn. The move event may cause the customer to either continue with their existing insurer, choose to churn partially (only changing home insurance) or churn completely. Figure 5.2.1 depicts partial churners (*A*) and complete churners (*B)* as subsets of the set *O*. Further, it is also true that in this case *B* is a subset of *A.*

$$B \subset A \subset O$$

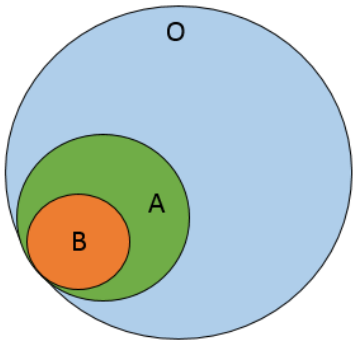


Figure 5.2. 1 The data used in Experiments 1-3

In order to appropriately capture these different layers of customer behavior, altogether three experiments are performed. They are defined as follows:

Experiment 1: $P(A|O)$

Which characteristics and preferences significantly affect a moving customer choosing to terminate their existing home insurance in favor of a competitor?

This is conducted by regressing the binary variable *home_churn* on our regressors depicting characteristics and preferences, with the dataset *O* in a logistic regression model.

Experiment 2: $P\ (B|A)$

Which characteristics and preferences significantly affect a partially churned customer's probability of churning completely?

This is conducted by regressing the binary variable *cust_churned* on our regressors, with the dataset *A* in a logistic regression model.

Experiment 3: $P\ (B|O)$

Which characteristics and preferences significantly affect a moving customer choosing to completely churn?

This is conducted by regressing the binary variable *cust_churned* on our regressors, with the dataset *O* in a logistic regression model.

5.2.1. Results: Experiment 1

Let the output yielded by Experiment 1 be known as *Model 1*. The first regression was performed with all relevant variables (21) as regressors. Model speficiation is conducted by variance analysis, removing non-significant variables that do not cause a decrease in residual deviance. The *anova* –table is available in the Appendix. After specification, the regression was performed again with 11 significant predictors. The results of the specified model are presented below, in Figure 5.2.2

```
> summary(homedata1)

Call:
glm(formula = home_churn_all ~ ., family = binomial(link = "logit"),
    data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2758  -0.7402  -0.6095  -0.4448   2.3872

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.500e+00  1.088e-01 -13.786  < 2e-16 ***
duration10+ yrs                 -3.142e-01  6.680e-02  -4.704 2.55e-06 ***
duration2 yrs                    7.554e-03  6.392e-02   0.118 0.905918
duration3-5 yrs                 -1.280e-01  5.688e-02  -2.250 0.024466 *
duration5-10 yrs                -1.160e-01  6.269e-02  -1.851 0.064149 .
days_since_last_CX_received     -2.952e-05  2.453e-05  -1.204 0.228763
age_group39-50                  -2.301e-01  5.103e-02  -4.508 6.54e-06 ***
age_group51-64                  -4.738e-01  5.811e-02  -8.153 3.55e-16 ***
age_group65+                    -6.965e-01  7.804e-02  -8.925  < 2e-16 ***
age_groupunder 27               -6.759e-03  4.249e-02  -0.159 0.873614
Ctc_emailYES                     1.961e-01  6.770e-02   2.897 0.003766 **
eDMs_last_90d                   -2.058e-02  1.069e-02  -1.925 0.054211 .
Gndr_CdMALE                     -6.009e-02  3.284e-02  -1.830 0.067271 .
nr_lob_personal_90_days_ago1    -6.556e-02  4.236e-02  -1.548 0.121721
nr_lob_personal_90_days_ago2    -2.723e-01  4.958e-02  -5.493 3.95e-08 ***
nr_lob_personal_90_days_ago3+   -3.850e-01  6.751e-02  -5.702 1.18e-08 ***
nr_lob_property_90_days_agoSINGLE  3.941e-01  4.382e-02   8.995  < 2e-16 ***
number_of_inbound_calls_90d     -6.435e-02  2.575e-02  -2.500 0.012430 *
Web_visits_last_45_days          4.423e-02  2.174e-02   2.034 0.041925 *
payment_plan2                    1.757e-01  6.671e-02   2.634 0.008435 **
payment_plan3                    1.424e-01  7.309e-02   1.948 0.051416 .
payment_plan4                    1.086e-01  5.335e-02   2.035 0.041838 *
payment_plan6                    2.315e-01  6.791e-02   3.409 0.000652 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24374  on 24229  degrees of freedom
Residual deviance: 23678  on 24207  degrees of freedom
AIC: 23724

Number of Fisher Scoring iterations: 4
```

Figure 5.2. 2 - The results of a specified Model 1.

The estimated coefficient for *nr_lob_property_90_days_ago_SINGLE* (0.3941) indicates direct proportionality with the home-churn probability. This implies – at the 99.9 % level - that customers with more than one policy in the property line of business are less likely to churn than those with a single home insurance.[4] There are two logical explanations to this:

1. Changing your home insurance from one insurer to another is quick. It is reasonable to assume that customers with multiple homes (and/or pets) won't want to go to the trouble

[4] An example of having more than one property item is when a customer has insured their summer home (and/or pet) in addition to their home.

quite so lightly as the amount of time and effort required increases with the number of policies.

2. The amount of property insured can be viewed as an indicator for customer wealth. Those with fewer belongings to insure might be more price sensitive than those with many, and can therefore be more susceptible to move their home insurance if a cheaper alternative is presented by a competitor.

Explanation 1 also holds true for the number of personal line products a customer has. Customers with two personal line products see their log-probability of churn drop by a factor 0.273 and those with three or more by 0.385 (at the 99.9 % level), when compared to customers with no personal line products. Where a high number of property products can be reasonably argued to imply wealth, it is not as evident with personal insurance. There is, however, another explanation for personal products' positive effect on retention. – Explanation 3.

3. Most personal line products, such as health and life insurance, require a thorough risk selection process before the contract may be started, in order to understand the potential customer's health history. The process may result in certain pre-existing conditions being ruled out of the product's coverage. This means that moving a health or life insurance from one insurer to another will quite often result in poorer coverage in the new product.

In summation, Explanation 1 implies that the more policies a customer has with their insurer, the less likely they are to churn their home insurance when it's time to move homes. Explanation 2 states that a high number of property products implies a low probability for home insurance churn whereas Explanation 3 states the same for products in the personal line. It is noteworthy that the amount of motor products in itself is not significant at the 10 %-level. [5]

Older customers, as well as customers with a longer tenure (captured by the *duration* –variable) seem to have a lower churn probability. An individual with a customership ten years or longer sees their log-odds of churning reduced by 0.3142 when compared to a first-year customer at the 99.9 % level. Similarly, a customer over the age of 65 presents log-odds of churning reduced by 0.695, when compared to one under the age of 38.

[5] Possible explanation: the main motor product (third party liability) is a regulated, statutory insurance. It is much easier to change motor insurance providers online than it is property or personal.

Premium is paid by the customer to the insurer in advance. The customer may choose to either pay the annual premium all at once, or break it down into smaller installments of either 2, 3, 4 or 6. Customers paying in six installments present a churn probability increase of 0.2315 in comparison to those paying in one sitting (at the 99.9 % level). Preferring to pay in many installments over just one may indicate price sensitivity, similarly to the way the customer's number of policies did.

The estimates indicate positive relationship between retention and customer communications. Telephone calls to the insurer's customer service as well as digitally sent communications to the customer decrease the probability of home insurance churn (at 95 % and 90 % levels, respectively). Web visits done by the customer over 45 days prior to the reported address change, however, increase the probability of home insurance churn at the 95 % level. Finally, females are more likely to move their home insurance than men, as evidenced by significance of the *Gndr_Cd* variable at the 90 % level.

5.2.2. Results: Experiment 2

Let the output produced by Experiment 2 be known as Model 2. Model 2 estimates the magnitude and significance of variables causing customers who change their home insurance to churn completely. It is important to note that the process of churning completely is not necessarily done in one sitting - changing providers may take days, even weeks. As the address change forces a customer to make the churn/stay –decision, it is not unreasonable to assume they would make that decision first, before beginning the process of assessing other policies they might have active. The fact that a moving customer choosing to change home insurance providers does not yet necessarily know whether they will churn completely or not lends credibility to the possibility that the results in Models 1, 2 and 3 may differ.

The data used in this model is the subset *A* (Figure 5.2.1), containing 4 891 observations. The subset is obtained by sampling from the set *O*, only selecting rows where the binary variable *home_churn_all* takes the value "1". The first step is to again conduct the regression with all 21 variables. The model is specified by way of variance analysis - as in Experiment 1 – to only include significant variables. The specified model included 12 variables, results of which are

presented in Figure 5.2.3. The table of variance analysis is for the unspecified model is available in the Appendix.

```
> homedata2 <- glm(cust_churned ~.,family=binomial(link='logit'),data=data
> summary(homedata2)

Call:
glm(formula = cust_churned ~ ., family = binomial(link = "logit"),
    data = data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2759  -0.9445   0.2785   0.6247   2.6977

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   2.406e+00  1.792e-01  13.429  < 2e-16 ***
duration10+ yrs              -8.551e-01  1.501e-01  -5.699 1.21e-08 ***
duration2 yrs                -1.648e-01  1.415e-01  -1.164 0.244252
duration3-5 yrs              -5.331e-01  1.259e-01  -4.235 2.29e-05 ***
duration5-10 yrs             -6.893e-01  1.389e-01  -4.962 6.98e-07 ***
days_since_last_CX_received   5.492e-05  5.216e-05   1.053 0.292304
age_group39-50                1.417e-02  1.134e-01   0.125 0.900514
age_group51-64                3.566e-02  1.310e-01   0.272 0.785456
age_group65+                  8.075e-01  1.741e-01   4.638 3.51e-06 ***
age_groupunder 27            -2.113e-01  9.088e-02  -2.325 0.020085 *
CX_forms_received_last_90d    1.729e-02  8.212e-02   0.211 0.833214
eDMs_last_90d                -1.129e-02  2.341e-02  -0.482 0.629533
Gndr_CdMALE                  -9.678e-02  7.467e-02  -1.296 0.194980
nr_lob_motor_90_days_ago1-2  -1.910e+00  8.554e-02 -22.330  < 2e-16 ***
nr_lob_motor_90_days_ago3-4  -2.282e+00  1.558e-01 -14.642  < 2e-16 ***
nr_lob_motor_90_days_ago5+   -2.620e+00  2.770e-01  -9.461  < 2e-16 ***
nr_lob_personal_90_days_ago1 -1.782e+00  9.292e-02 -19.172  < 2e-16 ***
nr_lob_personal_90_days_ago2 -1.784e+00  1.094e-01 -16.307  < 2e-16 ***
nr_lob_personal_90_days_ago3+ -1.997e+00  1.666e-01 -11.988  < 2e-16 ***
number_of_inbound_calls_90d  -2.261e-01  6.797e-02  -3.327 0.000879 ***
payment_plan2                -1.225e-01  1.466e-01  -0.835 0.403599
payment_plan3                 8.979e-03  1.630e-01   0.055 0.956078
payment_plan4                -2.467e-01  1.187e-01  -2.078 0.037702 *
payment_plan6                 2.710e-02  1.510e-01   0.179 0.857598
tm_contactedYES              -6.155e-01  1.833e-01  -3.358 0.000786 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6780.4  on 4890  degrees of freedom
Residual deviance: 4882.0  on 4866  degrees of freedom
AIC: 4932

Number of Fisher Scoring iterations: 4
```

Figure 5.2. 3 - The results of a specified Model 2

Model 2 estimates the effect of owning multiple policies to have a similarly positive effect on retention as did Model 1, with the addition of the presence of products in the motor line of business showing significance at the 99.9 % level. The variable depicting customership tenure behaves in a similar way to Model 1, but *age_group* presents an interesting result. In Model 1, older age implied a lower probability of home insurance churn. However according to Model 2, out of customers who change home insurance providers, those over the age of 65 have significantly increased log-odds of churning completely (0.8068 at the 99.9 % level) in comparison to the age 28-38 –segment. Further, customers under the age of 27 have higher

odds of retention (given a churned home insurance), than any other age group (-0.2115 at the 95 % level). These findings can be interpreted as older customers preferring to insure everything with one provider, whereas the young segment does not mind insurance cover scattered across multiple insurers.

A notable addition to the regressors in Model 2 when compared to Model 1 is the significance of the variable *tm_contacted*. Out of customers who changed home insurance providers, those who had received a telephone call from their original insurer no longer than 90 days prior were far less likely to churn completely than those who had not (*tm_contactedYES:* -0.6155 at 99.9 % significance level). Gender of the customer as well as the preference of payment installations lose significance going from Model 1 to Model 2.

5.2.3 Results: Experiment 3

Let the output yielded by Experiment 3 be known as Model 3. Model 3 estimates the magnitudes and significances of the effects variables have on complete customer churn. Our dataset is the full set *O*, out of whom set *B* are the customers for whom the binary variable *cust_churned* takes the value "1". Model 1 focused on what causes movers to change home insurance providers and Model 2 looked at what made home insurance churners choose to churn completely. Model 3 can therefore be seen as a validation experiment for the relationship between home insurance churn and complete churn; if the results are significantly different, we want to understand why. Dataset *O* consists of 24 230 observations with 24 columns of variables. Specification is performed by variance analysis and non-significant variables are removed. The estimation results of a specified Model 3, with 9 significant regressors, are presented in Figure 5.2.4.

```
> summary(homedata3x)

Call:
glm(formula = cust_churned ~ ., family = binomial(link = "logit"),
    data = train3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8737  -0.4890  -0.3595  -0.2232   3.4655

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.20683    0.13014  -9.274  < 2e-16 ***
duration10+ yrs                 -0.58141    0.10730  -5.419 6.01e-08 ***
duration2 yrs                    0.01839    0.09064   0.203 0.839193
duration3-5 yrs                 -0.25004    0.08249  -3.031 0.002436 **
duration5-10 yrs                -0.44437    0.09964  -4.460 8.20e-06 ***
age_group39-50                  -0.24461    0.09173  -2.667 0.007660 **
age_group51-64                  -0.50025    0.10338  -4.839 1.30e-06 ***
age_group65+                    -0.60181    0.11212  -5.367 7.99e-08 ***
age_groupunder 27               -0.06683    0.06655  -1.004 0.315292
Gndr_CdMALE                     -0.12804    0.05563  -2.302 0.021358 *
nr_lob_motor_90_days_ago1-2     -0.92425    0.07129 -12.965  < 2e-16 ***
nr_lob_motor_90_days_ago3-4     -1.01705    0.14527  -7.001 2.54e-12 ***
nr_lob_motor_90_days_ago5+      -1.70421    0.31142  -5.472 4.44e-08 ***
nr_lob_personal_90_days_ago1    -0.90768    0.08054 -11.270  < 2e-16 ***
nr_lob_personal_90_days_ago2    -1.04527    0.09452 -11.059  < 2e-16 ***
nr_lob_personal_90_days_ago3+   -1.43791    0.16518  -8.705  < 2e-16 ***
nr_lob_property_90_days_agoSINGLE  0.14735  0.07614   1.935 0.052960 .
number_of_inbound_calls_90d     -0.22186    0.05756  -3.855 0.000116 ***
payment_plan2                    0.22492    0.09893   2.273 0.023001 *
payment_plan3                    0.24112    0.10608   2.273 0.023029 *
payment_plan4                    0.16585    0.08224   2.017 0.043726 *
payment_plan6                    0.27474    0.11913   2.306 0.021097 *
tm_contactedYES                 -0.35349    0.15232  -2.321 0.020302 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11013.8  on 16960  degrees of freedom
Residual deviance:  9883.3  on 16938  degrees of freedom
AIC: 9929.3

Number of Fisher Scoring iterations: 6
```

Figure 5.2. 4 The results of a specified Model 3

Age and customership tenure affect full churn in a similar manner to Model 1, with older age and longer tenure lowering the churn probability. The variables with the largest estimates are the variables indicating insurance portfolio size. Customers with multiple products from the personal and motor lines of business have substantially better log-odds of survival, than those with none. A telephone contact – initiated either by the customer or the insurer – results in an improved probability of being retained.

5.3. Interpretation and discussion of results

This section combines the results of the three models and offers interpretation. Statistics depicting accuracy and predictive power are offered along with discussion thereof.

5.3.1 Model evaluation

	MODEL 1	**MODEL 2**	**MODEL 3**
McFadden R^2	0.029	0.280	0.103
AUC	0.601	0.793	0.722
χ^2	696.24***	1898.4***	1130.5***

5.3. 1 Statistics depicting goodness of fit and accuracy of Models 1, 2 and 3

The results described in section 5.2 describe the intensities and significances of the relationships between churn and customer variables. What has not yet been discussed, are the answers to the following questions:

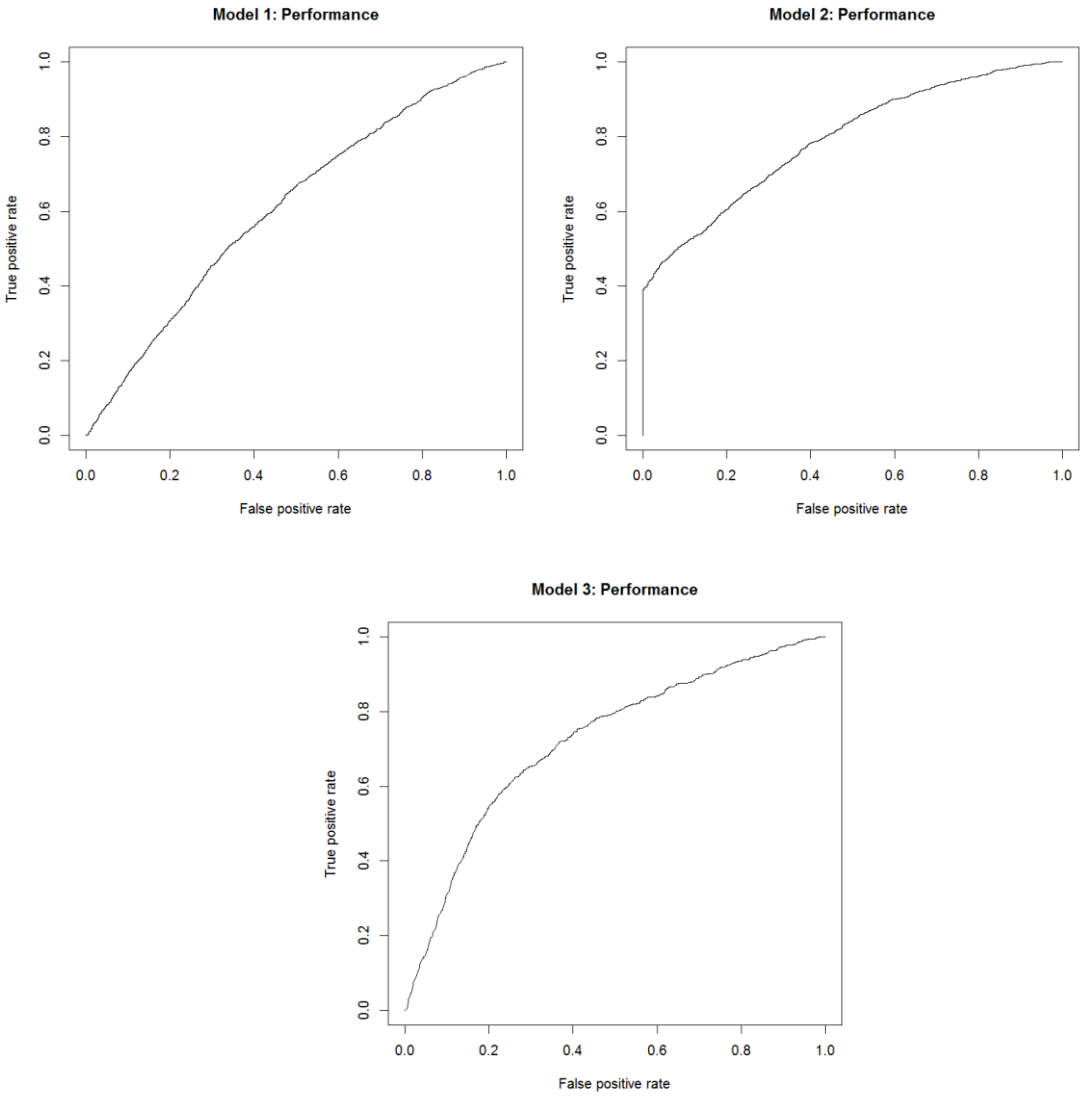
1. *How much* of the variance in our churn variables does our data explain?
2. *How accurately* do our models predict the correct outcome regarding churn?

Figure 5.3.1 offers help. The first row describes model accuracy via a statistic denoted here as *McFadden* R^2. McFadden (1974) proposed that a traditional R^2 index is not as well behaved for models using maximum likelihood estimation – such as our logit models – as it is for ordinary least squares (OLS) estimation. The statistic used here is a so-called pseudo-R^2, meant to be interpreted in the same way as the traditional measure. A higher value of McFadden's R^2 indicates a model that fits, i.e. explains the variation, better than a model with a lower value (Hemmert et al 2016).

Though McFadden (1974) argued his pseudo-R^2 –measure should always be expected to be lower in value than its more traditional counterpart, it can be said that none of the three models fit their respective data very well. Model 2, whose data consists only of customers having already churned their home insurance, displays the best indication of accuracy at 28 %.

The middle row provides another way by which to evaluate the predictive power of the model. The *area under curve* (AUC) statistic describes how accurately a model can predict a binary outcome correctly. (Allison 2014). Figure 5.3.2 presents the curves in question, known as the

receiver operating characteristic –curves. The curve plots together the true and false positive rates – a curve closer to the top left of the box implies greater accuracy. The AUC-statistic provides insight very similar to McFadden's R^2. The model used for home insurance churn (Model 1) does not fit its data particularly well (i.e. does not have great predictive power). Model 3 does slightly better in identifying complete churners from the mass of movers, but the most accurate model is Model 2, predicting complete churn correctly in close to 80 % of cases, where home insurance has been churned.



5.3. 2 The ROC curves of Models 1, 2 and 3

The bottom row displays the resulting statistic for conducting a Likelihood Ratio Test on each model. The high value of the χ^2 –statistics means that we reject the null hypothesis of the set of coefficients not being significantly different from zero at the 99.9 % level (Madden et al 1999). In other words, the model is deemed to fit the data.

Of importance is that these findings do not undermine the results obtained in section 5.2. Rather, it can be said there exists a variable not captured by the data that significantly affects the decision to churn. An explanation is proposed in the following section.

5.3.2 Interpreting the results

Figure 5.3.3 summarizes key columns from the regression results of Models 1, 2 and 3, while reporting the exponent of the estimated coefficients. This means that for each variable, the actual estimates for the probability effects are reported respectively under each model in the column "*exp(estimate)*".

From Figure 5.3.3 it can be seen that the most influential variables affecting both home insurance- and complete churn among moving customers are the length of customership tenure (*duration*) and the presence of personal insurance products. From Model 1, we see that a moving customer most likely to change home insurance providers is young, with only one home insured, paying premium in multiple installments.

Differing greatly from Models 1 and 3, Model 2 suggests older customers are actually significantly more likely to churn completely, given a changed home insurance. A telephone contact made by the insurer has no effect on home insurance churn, but is a significant mitigator of complete customer churn. Electronic communications seem to have no significant effects on churn. In general a smaller insurance portfolio and paying in multiple installments indicate a higher churn probability. On the other hand, recent telephone communications and longer customership tenure affect retention in a positive way.

VARIABLE	MODEL 1 *home churn on full data* exp(estimate)	significance	MODEL 2 *full churn on home churners* exp(estimate)	significance	MODEL 3 *full churn on full data* exp(estimate)	significance
duration 10+	0.730	***	0.425	***	0.559	***
duration 2	1.008		0.848		1.019	
duration 3-5	0.880	*	0.587	***	0.779	**
duration 5-10	0.890	'	0.502	***	0.641	***
age 39-50	0.794	***	1.014		0.783	**
age 51-64	0.623	***	1.036		0.607	***
age 65+	0.499	***	2.242	***	0.548	***
age under27	0.509		0.810	*	0.935	
Ctc_email YES	1.217	**	-		-	
CX_forms_90d	-		1.017		-	
eDMs 90d	0.980	'	0.989		-	
Gndr_Cd MALE	0.942	'	0.908		0.880	*
nr_motor 1-2	-		0.148	***	0.397	***
nr_motor 3-4	-		0.102	***	0.362	***
nr_motor 5+	-		0.073	***	0.182	***
nr_personal 1	0.937		0.168	***	0.403	***
nr_personal 2	0.762	***	0.168	***	0.350	***
nr_personal 3	0.680	***	0.136	***	0.237	***
nr_property SINGLE	1.483	***	-		1.158	'
number inbound 90d	0.938	*	0.798	***	0.801	***
tm_contacted YES	-		0.540	***	0.703	*
Web_visits_45d	1.045	*	-			
payment 2	1.192	**	0.885		1.252	*
payment 3	1.153	'	1.009		1.273	*
payment 4	1.115	*	0.781	*	1.181	*
payment 6	1.260	***	1.027		1.317	*

*Significance codes: *** = 0.001 // ** = 0.01 // * = 0.05 // ' = 0.1*

5.3. 3 5.3. 3 Actual factors by which churn probabilities are affected + significances for Models 1, 2 and 3

From the previous chapter we obtained that however insightful these findings are, they explain only a fraction of the variance in our churn variables. It is implied there should exist such a variable (or variables) not captured by our data, that would help explain the decision to churn in a significant way. The experiments conducted in 5.2 differ from previous work by not having price or competitor-related variables in our data. In Madden et al (1999) the most significant reason for churn as reported by customers was the price of internet access. It is reasonable to

assume this fact holds also for insurance contracts: the price offered by one's original home insurance provider as well as the prices of competitors should affect the decision regarding home insurance (the choice problem in Model 1) significantly. As Models 2 and 3 fit their respective data better than Model 1, it can be interpreted that price and competitor activity affect the decision regarding other products in a customer's insurance portfolio less than they do home insurance. Another argument for economic factors being crucially important in the churn choice problem is the significance of wealth-indicating variables in the models. Both the preference of paying premium in one installment and the size of insurance portfolio indicate wealth – i.e. less pressure for said customer to churn given a more affordable alternative.

I assume differences in price between providers to be the single most important driver of home insurance churn at the time of move, especially in lower income individuals. However variables used in our data, depicting non-price elements, play a significant role in potentially saving a larger customership even if the home insurance is initially lost.

6 Conclusions

This thesis proposed the research question: "*what are such customer characteristics and preferences that indicate an increased probability for churning at the time of a move*?" This question was broken down into three in Chapter 5. We found that customers indicating lower income, smaller insurance portfolio and a shorter customership tenure were more likely to change home insurances. We also found that older customers were far more likely to change all insurances, given a decision to change the home policy. Further, we found that a long customership, a larger insurance portfolio – especially in the personal line of business – as well as contact with the original insurer are crucial in preventing full customer churn.

There were one million moving individuals in Finland in 2016, per Posti (2017a). We can approximate that for an insurer, one in five customers is a mover at some point during the year. The results discussed here offer insight into ways of segmenting the mass of moving customers based on the estimated coefficients of the variables. Presented are some examples of said segmentation.

The results suggest that a telephone contact by the insurer is not a significant mitigator of home insurance churn. Therefore it is not of paramount importance to contact moving customers as early as possible. Rather, recognizing instances where a home insurance has been churned and treating those customers with a call should yield better results.

The variable denoting customership tenure has significant and large coefficients and therefore requires discussion. A customership is *active* as long as at least one insurance policy is active for that customer. It should be in the insurers' interest to care for and prolong customers in the risk-area, i.e. the first two years of customership, by focusing communications on them. Some segmentation should also be done based on payment preference – more focus on customers paying in multiple installments.

The most crucial improver of movers' retention, as reported by all three models, is the number of personal insurance policies the customer subscribes to. Whereas motor insurance is mostly statutory (discussed in 5.2) and home insurance is *de-facto* statutory, personal insurance is optional. The best way for an insurer to combat churn would be to upsell their customers with

optional personal insurance in every contact. It is somewhat abstract to offer this as a treatment to churn as it is more a long-term commitment by an insurer, rather than a quickly applicable bandage. It is nonetheless to be assumed, that the larger the share of movers with more than one personal product, the lower the churn rate.

This descriptive econometric research may be followed up by a causal study, testing treatments such as the ones proposed above, assessing the applicability of these findings.

As discussed in 5.3, this thesis accounts for economic variables only via proxies. The competitive nature of subscription-based businesses as well as the low *McFadden* R^2 –scores of the models indicate a presence of heavy price sensitivity. A survey-based study may be more appropriate for a research problem such as this one, not only to determine customers' attitude towards price but to capture service attributes of more than one insurer, as McFadden originally intended.

A final notion for future research. The models constructed in this thesis treats all customers as equals in the insurer's eyes. To understand the true effect of churn on an insurer's result this churn model may be complemented with a customer lifetime value –calculation. A CLV-model would score a customer based on portfolio size and duration (possibly also profitability). Combining such a model with the churn model constructed in this thesis would help the insurer in making the decision between targeting a customer with a lower CLV-score but with a higher churn probability and a high value customer with a slightly lower churn score.

6 References

Ahn et al. (2006). "Customer churn analysis", *Telecommunications Policy 30; 2006; 552-568.*

Allison, Paul D. (2014). "Measures of Fit for Logistic Regression", *Statistical Horizons LLC and the University of Pennsylvania, Paper 1485-2014, SAS Global Forum.*

Ben Akiva, M. & Lerman, S. (1985). "Discrete Choice Analysis: Theory and Application to Travel Demand", *MIT Press.*

Cox, D. R. (1958). "The Regression Analysis of Binary Sequences", *Journal of the Royal Statistical Society. Series B (Methodological), Vol 20, No 2. (1958), pp 215-242.*

Finance Finland (2017). "Finnish Insurance in 2016 – Financial overview of Finnish insurance companies", *Finance Finland FFI 31.05.2017,* <available online>

Hartley, James E. (1996). "Retrospectives: The Origins of the Representative Agent", *Journal of Economic Perspectives – Vol. 10, Number 2, Spring 1996, pp. 169-177.*

Hemmert, Giselmar A.J. et al. (2016). "Log-likelihood based Pseudo-R^2 in Logistic Regression" *Deriving Sample-sensitive Benchmarks. First published online March 18, 2016.*

Hillson, D. & Murray-Webster, R. (2007). "Understanding and Managing Risk Attitude", *Gower Publishing Ltd., 2007*

Insurance Companies Act (2008). "Insurance Companies Act 2008/521 §15", *18.07.2008, Helsinki.* *<available electronically>*

Insurance Contracts Act (1994). "Insurance Contracts Act 1994/543 §7, §12, §16, §17a", *28.06.1994, Helsinki.* *<available electronically>*

Jaccard, J. & Jacoby, J. (2010). "Theory construction and model-building skills: a practical guide for social scientists", *The Guilford Press, New York.*

Jevons, William Stanley (1871). "The Theory of Political Economy", *James R. Newman, ed., The World of Mathematics, Vol. 2, Part IV, 1956.*

Kim, Hee-Su (2004). "Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market", *Telecommunications Policy, Vol 28, Issues 9-10, Oct-Nov 2004, pp 751-765.*

Kunreuther, Howard et al. (2013). "Insurance & Behavioral Economics", *Cambridge University Press, 2013.*

Madden, G., Savage, S. & Coble-Neal, G. (1999). "Subscriber churn in the Australian ISP market", *Information Economics and Policy, Vol 11, Issue 2, pp 195-201.*

Marshack, J & Block, H.D. (1960). "Random Orderings and Stochastic Theories of Responses", *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, Chapter 10, pp. 97-*

McFadden, Daniel (1974). "Conditional Logit Analysis of Qualitative Choice Behavior", *Frontiers in Econometrics, Chapter Four, pp 105-142.*

McFadden, Daniel (1981). "Econometric Models for Probabilistic Choice Among Products", *The Journal of Business, Vol 53, No. 3, Part 2, pp S13-S29.*

McFadden, Daniel (2000). "Economic Choices", *Nobel Prize Lecture 2000, publ. The American Economic Review, Vol. 91, No. 3. (Jun., 2001), pp. 351-378.*

Pareto, Vilfredo (1906). "Manual of Political Economy: A Variorum Translation and Critical Edition", *Oxford University Press 2014.*

Posti (2016). "Postin muuttajatuotteet", *Marketing services, Posti, 19.12.2016.* <available electronically>

Posti (2017a). "Miljoona muuttoa vuodessa", *Study conducted by Posti and Statistics Finland.*

Posti (2017b). "Miten Suomi muuttaa", *Muuttajatilastoja Suomesta*

Small, Kenneth A. (2005). “Fundamentals of Economic Demand Modeling: Lessons From Travel Demand Analysis”, *Decision-Based Design: Making Effective Decisions in Product and Systems Design, Chapter 9.*

Varian, Hal (1987). “Intermediate Microeconomics – A Modern Approach”, *W. W. Norton & Company, New York, 1987.*

Vaughan, Emmett J. (1996). “Fundamentals of Risk and Insurance”, *Seventh edition, John Wiley & Sons, Inc., 1996*

Appendix

The full list of variables before manipulation:

```
[1] "duration"                          "days_since_last_CX_received"
 [3] "days_since_last_CX_answered"       "age_group"
 [5] "Ctc_email"                         "Ctc_phone"
 [7] "Customer_Commented_Eff"            "Customer_Commented_Imp"
 [9] "CX_forms_answered_last_90d"        "CX_forms_received_last_90d"
[11] "DMs_last_90d"                      "eDMs_last_90d"
[13] "Gndr_Cd"                           "Ins_Polcy_Dlv_Cd"
[15] "nr_lob_motor_90_days_ago"          "nr_lob_personal_90_days_ago"
[17] "nr_lob_property_90_days_ago"       "number_of_inbound_calls_90d"
[19] "num_active_objects_90_days_ago"    "times_tm_contacted_last_90d"
[21] "Web_visits_last_45_days"           "cust_churned"
[23] "home_churn_all"                    "payment_plan"
[25] "had_personal_90d"                  "had_motor_90d"
[27] "tm_contacted"
```

The *anova* –tables for Models 1, 2 and 3:

Model 1:

```
> anova(homedata1,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: home_churn_all

Terms added sequentially (first to last)


                           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                       24229      24374
duration                    4  303.996     24225      24070 < 2.2e-16 ***
days_since_last_CX_received 1    7.636     24224      24062  0.005721 **
days_since_last_CX_answered 1   17.504     24223      24045 2.866e-05 ***
age_group                   4  183.619     24219      23861 < 2.2e-16 ***
Ctc_email                   1    3.473     24218      23858  0.062368 .
Ctc_phone                   1    0.139     24217      23857  0.709034
Customer_Commented_Eff      1    1.261     24216      23856  0.261524
Customer_Commented_Imp      1    0.284     24215      23856  0.594328
CX_forms_answered_last_90d  1    5.708     24214      23850  0.016891 *
CX_forms_received_last_90d  1    0.313     24213      23850  0.576050
DMs_last_90d                1    4.608     24212      23845  0.031829 *
eDMs_last_90d               1    4.389     24211      23841  0.036165 *
Gndr_Cd                     1    3.079     24210      23838  0.079292 .
Ins_Polcy_Dlv_Cd            1    5.954     24209      23832  0.014681 *
nr_lob_motor_90_days_ago    3    3.374     24206      23828  0.337503
nr_lob_personal_90_days_ago 3   61.765     24203      23767 2.467e-13 ***
nr_lob_property_90_days_ago 1   81.128     24202      23686 < 2.2e-16 ***
number_of_inbound_calls_90d 1    5.033     24201      23681  0.024868 *
times_tm_contacted_last_90d 1    0.033     24200      23681  0.856031
Web_visits_last_45_days     1    3.898     24199      23677  0.048357 *
payment_plan                4   13.770     24195      23663  0.008067 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Model 2:

```
> anova(homedata2,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cust_churned

Terms added sequentially (first to last)

                            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                         4890     6780.4
duration                     4   245.61      4886     6534.8 < 2.2e-16 ***
days_since_last_CX_received  1   152.53      4885     6382.2 < 2.2e-16 ***
days_since_last_CX_answered  1     8.80      4884     6373.4 0.0030152 **
age_group                    4    59.83      4880     6313.6 3.156e-12 ***
Ctc_email                    1    11.75      4879     6301.9 0.0006084 ***
Ctc_phone                    1    14.59      4878     6287.3 0.0001340 ***
Customer_Commented_Eff       1     0.00      4877     6287.3 0.9824860
Customer_Commented_Imp       1     0.34      4876     6286.9 0.5587700
CX_forms_answered_last_90d   1     1.64      4875     6285.3 0.2001897
CX_forms_received_last_90d   1    38.07      4874     6247.2 6.812e-10 ***
DMs_last_90d                 1     4.30      4873     6242.9 0.0382014 *
eDMs_last_90d                1    28.72      4872     6214.2 8.371e-08 ***
Gndr_Cd                      1    41.06      4871     6173.1 1.473e-10 ***
Ins_Polcy_Dlv_Cd             1     1.02      4870     6172.1 0.3135335
nr_lob_motor_90_days_ago     3   659.51      4867     5512.6 < 2.2e-16 ***
nr_lob_personal_90_days_ago  3   604.91      4864     4907.7 < 2.2e-16 ***
nr_lob_property_90_days_ago  1     9.29      4863     4898.4 0.0023065 **
number_of_inbound_calls_90d  1    10.08      4862     4888.3 0.0015011 **
times_tm_contacted_last_90d  1    10.34      4861     4878.0 0.0013053 **
Web_visits_last_45_days      1     2.08      4860     4875.9 0.1489059
payment_plan                 4    11.26      4856     4864.7 0.0238168 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 3:

```
> anova(homedata3x,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cust_churned

Terms added sequentially (first to last)

                            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                        16960    11013.8
duration                     4   360.93     16956    10652.9 < 2.2e-16 ***
days_since_last_CX_received  1    33.72     16955    10619.2 6.370e-09 ***
days_since_last_CX_answered  1     9.52     16954    10609.7 0.0020280 **
age_group                    4    57.53     16950    10552.1 9.555e-12 ***
Ctc_email                    1     0.07     16949    10552.1 0.7946699
Ctc_phone                    1     2.46     16948    10549.6 0.1171425
Customer_Commented_Eff       1     0.00     16947    10549.6 0.9606621
Customer_Commented_Imp       1     2.61     16946    10547.0 0.1060934
CX_forms_answered_last_90d   1     0.10     16945    10546.9 0.7498246
CX_forms_received_last_90d   1    24.73     16944    10522.2 6.596e-07 ***
DMs_last_90d                 1     4.94     16943    10517.2 0.0262542 *
eDMs_last_90d                1    18.75     16942    10498.5 1.494e-05 ***
Gndr_Cd                      1    24.53     16941    10474.0 7.332e-07 ***
Ins_Polcy_Dlv_Cd             1     1.88     16940    10472.1 0.1707068
nr_lob_motor_90_days_ago     3   265.68     16937    10206.4 < 2.2e-16 ***
nr_lob_personal_90_days_ago  3   308.79     16934     9897.6 < 2.2e-16 ***
nr_lob_property_90_days_ago  1     4.44     16933     9893.2 0.0351506 *
number_of_inbound_calls_90d  1    11.46     16932     9881.7 0.0007094 ***
times_tm_contacted_last_90d  1     5.14     16931     9876.6 0.0233734 *
Web_visits_last_45_days      1     0.23     16930     9876.3 0.6331509
payment_plan                 4     7.18     16926     9869.2 0.1269065
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### FINAL REGRESSIONS FOR THESIS

# The data is split into three, after which three experiments are
performed:
```

```
## 1. home_churn_all is regressed on regressors, full data.
# i.e. A ~ X [ P(A|X) ]

## 2. cust_churned is regressed on the regressors with data consisting only
of churners.
# i.e. B ~ A [ P(B|A) ]

## 3. Finally cust_churned is regressed on with the full data,
# i.e. B ~ X [ P(B|X) ]


########################### READ DATA + VARIABLE MANIPULATION ########

data <- read.table("gradu2812.csv",header=TRUE,sep=",",na.strings=c("?"))
colnames(data)
str(data)
length(data$duration)

# AGE

ageg <- data$age_group

table(ageg,data$home_churn_all)
plot(ageg,data$home_churn_all)

ageg <- as.character(ageg)
levels(ageg)
ageg[(ageg) == "0-17"] <- "under 27"
ageg[(ageg) == "18-20"] <- "under 27"
ageg[(ageg) == "21-26"] <- "under 27"
ageg[(ageg) == "27-32"] <- "27-38"
ageg[(ageg) == "33-38"] <- "27-38"
ageg[(ageg) == "39-44"] <- "39-50"
ageg[(ageg) == "45-50"] <- "39-50"
ageg[(ageg) == "51-56"] <- "51-64"
ageg[(ageg) == "57-65"] <- "51-64"
ageg[(ageg) == "65+"] <- "65+"
ageg <- as.factor(ageg)
levels(ageg)
data$age_group <- ageg

data$payment_plan <- as.factor(data$payment_plan)
levels(data$age_group)
str(data)

data$cust_churned <- (as.factor(data$cust_churned))
data$home_churn_all <- (as.factor(data$home_churn_all))
data$Gndr_Cd[data$Gndr_Cd == ""] <- "MALE"

data$motor90[(data$nr_lob_motor_90_days_ago) == 0] <- "0"
data$motor90[(data$nr_lob_motor_90_days_ago) == 1] <- "1-2"
data$motor90[(data$nr_lob_motor_90_days_ago) == 2] <- "1-2"
data$motor90[(data$nr_lob_motor_90_days_ago) == 3] <- "3-4"
data$motor90[(data$nr_lob_motor_90_days_ago) == 4] <- "3-4"
data$motor90[(data$nr_lob_motor_90_days_ago) > 4] <- "5+"
data$nr_lob_motor_90_days_ago <- as.factor(data$motor90)
levels(data$nr_lob_motor_90_days_ago)

data$personal90[(data$nr_lob_personal_90_days_ago) == 0] <- "0"
data$personal90[(data$nr_lob_personal_90_days_ago) == 1] <- "1"
data$personal90[(data$nr_lob_personal_90_days_ago) == 2] <- "2"
```

```
data$personal90[(data$nr_lob_personal_90_days_ago) > 2] <- "3+"
data$nr_lob_personal_90_days_ago <- as.factor(data$personal90)
levels(data$nr_lob_personal_90_days_ago)

data$property90[(data$nr_lob_property_90_days_ago) < 5 ] <- "SINGLE"
data$property90[(data$nr_lob_property_90_days_ago) > 4] <- "MULTIPLE"
data$nr_lob_property_90_days_ago <- as.factor(data$property90)
levels(data$nr_lob_property_90_days_ago)

# Categorize duration variable

data$duration_cat[(data$duration) < 366] <- "1 yr"
data$duration_cat[(data$duration) > 365 & (data$duration) < 731] <- "2 yrs"
data$duration_cat[(data$duration) > 730 & (data$duration) < 1826] <- "3-5
yrs"
data$duration_cat[(data$duration) > 1825 & (data$duration) < 3650] <- "5-10
yrs"
data$duration_cat[(data$duration) > 3649] <- "10+ yrs"
data$duration <- as.factor(data$duration_cat)
plot(table(data$duration))

# Drop irrelevant / redundant variables

data$home_churn <- NULL
data$duration_cat <- NULL
data$Effort_Score <- NULL
data$NPS_Score <- NULL
data$property90 <- NULL
data$personal90 <- NULL
data$motor90 <- NULL

## BEGIN TO SUBSET DATA INTO THREE (3) DATAFRAMES ACCORDING TO INTRO:
colnames(data)
str(data)

## 1. home_churn_all is regressed on regressors, full data.
# i.e. A ~ X [ P(A|X) ]

data1 <- data[c(1:13,14,15,16,17,18,20,21,23,24)]

## 70 of the sample size
smp_size1 <- floor(0.7 * nrow(data1))

## set the seed to make your partition reproductible
set.seed(123)
train_ind1 <- sample(seq_len(nrow(data1)), size = smp_size1)
train1 <- data1[train_ind1, ]
test1 <- data1[-train_ind1, ]

homedata1 <- glm(home_churn_all
~.,family=binomial(link='logit'),data=data1)
summary(homedata1)

pR2(homedata1)
anova(homedata1,test="Chisq")
fitted1 <- predict(homedata1, newdata=test1, type='response')
fitted1 <- ifelse(fitted1 > 0.5,1,0)
misClasificError1x <- mean(fitted1 != test1$home_churn_all)
print(paste('Accuracy',1-misClasificError1x))

library(ROCR)
```

```
p1 <- predict(homedata1, newdata=test1, type="response")
pr1 <- prediction(p1, test1$home_churn_all)
prf1 <- performance(pr1, measure = "tpr", x.measure = "fpr")
plot(prf1)

auc1 <- performance(pr1, measure = "auc")
auc1 <- auc1@y.values[[1]]
auc1




## 2. cust_churned is regressed on the regressors with data consisting only
of churners.
# i.e. B ~ A [ P(B|A) ]

data2 <- data[c(1:13,14,15,16,17,18,20,21,22,23,24)]
data2 <- subset(data2, home_churn_all == "1")
data2$home_churn_all <- NULL

length(data2$cust_churned)

## 70 of the sample size
smp_size2 <- floor(0.7 * nrow(data2))

## set the seed to make your partition reproductible
set.seed(123)
train_ind2 <- sample(seq_len(nrow(data2)), size = smp_size2)

train2 <- data2[train_ind2, ]
test2 <- data2[-train_ind2, ]

homedata2 <- glm(cust_churned ~.,family=binomial(link='logit'),data=data2)
summary(homedata2)
pR2(homedata2)
anova(homedata2,test="Chisq")

fitted2 <- predict(homedata2, newdata=test2, type='response')
fitted2 <- ifelse(fitted2 > 0.5,1,0)
misClasificError2x <- mean(fitted2 != test2$home_churn_all)
print(paste('Accuracy',1-misClasificError2x))

library(ROCR)
p2 <- predict(homedata2, newdata=test2, type="response")
pr2 <- prediction(p2, test2$home_churn_all)
prf2 <- performance(pr2, measure = "tpr", x.measure = "fpr")
plot(prf2)

auc2 <- performance(pr2, measure = "auc")
auc2 <- auc2@y.values[[1]]
auc2



## 3. Finally cust_churned is regressed on with the full data, SANS
home_churn_all
# i.e. B ~ X [ P(B|X) ]

data3 <- data[c(1:13,14,15,16,17,18,20,21,22,24)]
```

```
## 70 of the sample size
smp_size3 <- floor(0.7 * nrow(data3))

## set the seed to make your partition reproductible
set.seed(123)
train_ind3 <- sample(seq_len(nrow(data3)), size = smp_size3)

train3 <- data3[train_ind3, ]
test3 <- data3[-train_ind3, ]

homedata3x <- glm(cust_churned ~
.,family=binomial(link='logit'),data=train3)
summary(homedata3x)
anova(homedata3x,test="Chisq")
pR2(homedata3x)

fitted3 <- predict(homedata3x, newdata=test3, type='response')
fitted3 <- ifelse(fitted3 > 0.5,1,0)


misClasificError3x <- mean(fitted3 != test3$cust_churned)
print(paste('Accuracy',1-misClasificError3x))

library(ROCR)
p3 <- predict(homedata3x, newdata=test3, type="response")
pr3 <- prediction(p3, test3$cust_churned)
prf3 <- performance(pr3, measure = "tpr", x.measure = "fpr")
plot(prf3)

auc3 <- performance(pr3, measure = "auc")
auc3 <- auc3@y.values[[1]]
auc3

#########################################################################
## After running the experiments, it is found that not all variables######
# are significant. The anova table reveals the variables to be dropped.####
#########################################################################

colnames(data)
## 1. home_churn_all is regressed on regressors, full data.
# i.e. A ~ X [ P(A|X) ]

data1 <- data[c(1,2,4,5,12,13,16,17,18,21,23,24)]

## 70 of the sample size
smp_size1 <- floor(0.7 * nrow(data1))

## set the seed to make your partition reproductible
set.seed(123)
train_ind1 <- sample(seq_len(nrow(data1)), size = smp_size1)
train1 <- data1[train_ind1, ]
test1 <- data1[-train_ind1, ]

homedata1 <- glm(home_churn_all
~.,family=binomial(link='logit'),data=data1)
summary(homedata1)
exp(coef(homedata1))
pR2(homedata1)
anova(homedata1,test="Chisq")

library(ROCR)
```

```
p1 <- predict(homedata1, newdata=test1, type="response")
pr1 <- prediction(p1, test1$home_churn_all)
prf1 <- performance(pr1, measure = "tpr", x.measure = "fpr")
plot(prf1)

auc1 <- performance(pr1, measure = "auc")
auc1 <- auc1@y.values[[1]]
auc1



## 2. cust_churned is regressed on the regressors with data consisting only
of churners.
# i.e. B ~ A [ P(B|A) ]

colnames(data)
data2 <- data[c(1,2,4,10,12,13,15,16,17,18,20,22,23,24)]
data2 <- subset(data2, home_churn_all == "1")
data2$home_churn_all <- NULL

length(data2$cust_churned)

## 70 of the sample size
smp_size2 <- floor(0.7 * nrow(data2))

## set the seed to make your partition reproductible
set.seed(123)
train_ind2 <- sample(seq_len(nrow(data2)), size = smp_size2)

train2 <- data2[train_ind2, ]
test2 <- data2[-train_ind2, ]


homedata2 <- glm(cust_churned ~.,family=binomial(link='logit'),data=data2)
summary(homedata2)
pR2(homedata2)
exp(coef(homedata2))
anova(homedata2,test="Chisq")


library(ROCR)
p2 <- predict(homedata2, newdata=test2, type="response")
pr2 <- prediction(p2, test2$home_churn_all)
prf2 <- performance(pr2, measure = "tpr", x.measure = "fpr")
plot(prf2)

auc2 <- performance(pr2, measure = "auc")
auc2 <- auc2@y.values[[1]]
auc2



## 3. Finally cust_churned is regressed on with the full data, SANS
home_churn_all
# i.e. B ~ X [ P(B|X) ]

colnames(data)
data3 <- data[c(1,4,13,15,16,17,18,20,22,24)]

## 70 of the sample size
smp_size3 <- floor(0.7 * nrow(data3))
```

```
## set the seed to make your partition reproductible
set.seed(123)
train_ind3 <- sample(seq_len(nrow(data3)), size = smp_size3)

train3 <- data3[train_ind3, ]
test3 <- data3[-train_ind3, ]

homedata3x <- glm(cust_churned ~
.,family=binomial(link='logit'),data=train3)
summary(homedata3x)
exp(coef(homedata3x))
anova(homedata3x,test="Chisq")
pR2(homedata3x)


library(ROCR)
p3 <- predict(homedata3x, newdata=test3, type="response")
pr3 <- prediction(p3, test3$cust_churned)
prf3 <- performance(pr3, measure = "tpr", x.measure = "fpr")
plot(prf3)

auc3 <- performance(pr3, measure = "auc")
auc3 <- auc3@y.values[[1]]
auc3
```